http://journals.bmsu.ac.ir/ijtmgh
Int J Travel Med Glob Health. 2014;2(2):87-90
IJTMGH
International Journal of Travel Medicine and Global Health
Review Article Open Access

Data Mining: A Novel Outlook to Explore Knowledge in Health and Medical Sciences

Golbarg Mehrpoor[1], Mohammad Mehdi Azimzadeh[1,*], Amirhossein Monfared[2]

1. *School of Medicine, Alborz University of Medical Sciences, Karaj, Iran*
2. *Department of Industrial Intelligence Research Group, ACECR, Zanjan Branch, Zanjan, Iran*

[*]**Corresponding Author:** Mohammad Mehdi Azimzadeh, Software Engineering, School of Medicine, Alborz University of Medical Sciences, Karaj, Iran
Email: mehdiazimzadeh19@yahoo.com

Abstract

Today medical and Healthcare industry generate loads of diverse data about patients, disease diagnosis, prognosis, management, hospitals' resources, electronic patient health records, medical devices and etc. Using the most efficient processing and analyzing method for knowledge extraction is a key point to cost-saving in clinical decision making. Data mining, sometimes called data or knowledge discovery, is the process of analyzing data from different perspectives and summarizing it into useful information. In medicine, this process is distinct from that in other fields, because of heterogeneity and voluminosity of the data. Herein we reviewed some of published articles about application of data mining in several fields in medicine and healthcare.

Keywords: Medicine, Knowledge, Medical Informatics, Data Mining

Article History: **Received:** 18 Jun 2013 **Revised:** 1 Apr 2014 **Accepted:** 15 Apr 2014
Cite this article as: Mehrpoor G, Azimzadeh MM, Monfared A. Data Mining: A novel outlook to explore knowledge in medical and health sciences. Int J Travel Med Glob Health. 2014;2(2):87-90.

Introduction

Given the rapid and extensive development of science and technology, there is an urgent need to organize the information. In medical field, information system can considerably facilitate utilization of the information. Many sectors, nowadays, deal with electronic and non-electronic data, which can be converted into information and knowledge by employing different technologies.

Information technology was introduced in 1960 to convert preliminary data of information systems and databases. In those early days, systems such as database management system were only in charge of rapidly saving great deal of data. However, technological development did not stop there and the data were gradually converted into useful knowledge [1]. Data in the age of information is considered as a valuable asset. Hence, development of an organization is highly dependent on available information and the way such information is analyzed. Evidently, many organizations have initiated quality improvement programs under different titles. Such programs can be more fruitful if they enjoy better analysis of the stored data. It is notable, however, that interpretation of enormous deal of data is not an easy job. Data mining is one of the methods that help us in this regard.

Data mining is about inferring or extracting knowledge out of an enormous deal of data available. The term comes from "Gold Mining" which refers to extracting gold from mountains [2, 3]. The science of data mining employs intelligent techniques to extract knowledge out a set of data. Such knowledge is not achievable using conventional statistic techniques. The science has the same meaning with "knowledge discovery in database" or KDD [4]. Data mining was introduced in the late 1990s and with its rapid growth we can expect big changes in knowledge generation in the world.

A workshop of extracting knowledge out of databases was held from 1989 to 1994. The term "data mining" was first used by Fiaz et al. Since then, the idea has been seriously studied in statistics [5], and first data mining journal was published in 1996 [6].

Five key features of data mining:

Extracting, changing, and opening transaction data in a data storage system.

Saving and managing data in multi-dimensional information bank system.

Providing data access for commercial analyzers and experts of information.

Data analysis using applications;

Introducing a useful frame of data such as graphs or tables [7]. There are three general types of data mining including clustering, classification, and association rule discovery [9[.

A set of data, under cluster method, is clustered based on similarities and differences so that data in one cluster are similar and different from the data in other clusters. In general, clustering is the first step in data mining and put the data in the pertinent records for further analyses. The main clustering methods are classification, hierarchical, and density based [9].

Table 1. Differences between statistics and data mining [8]

Data mining does not need starting with a hypothesis	Statistic experts always start with a hypothesis
Data mining algorithm automatically build the relationships	Statistic experts must create relationship based on the hypotheses
Work with different types of data including numerical data	Numerical data are used
Data mining depends on accuracy and proper classification of data	Wrong and inappropriate data are detected throughout analysis
Results of data mining can be complex and need further statistical analyses to become understandable	The results can be interpreted and understandable for the managers

Classification is a learning process under supervision, which takes place in two stages. First, a set of data are employed to create a model of data. The model actually describes the concepts and features of a set of data on which the model is built. The second stage is to implement or utilize the created model of the data on the data including all features thereof [10].

There are different algorithms and methods proposed for classification; for instance, decision tree, Bayes classifier, SVM, classification by neural networks, and rule based classification [10].

Rules of association try to find elements and items in a set of data that commonly occur in the data so that an associations between occurrences of data is assumed [10].

Data mining applications

1. Retailing: some of classic applications of data mining are:
   - Determining customers' purchase pattern;
   - Analyzing market purchase portfolio;
   - Forecasting purchase through post (e-sale)
2. Insurance
   - Analyzing claims
   - Forecasting policy purchases
3. Banking:
   - Forecasting credit cards fraud
   - Determining fix number of customers
   - Determining popularity of credit cards based on social classes;
4. Space and space trips
   - Space information processing;
   - Space ship information processing;
   - Providing knowledge to make final decision to launch the space ship.

Data mining and cancer

Data mining nowadays is one of the best ways to diagnose and treat cancer as well as early diagnosis of it [11]. Artificial neural network (ANN) is one the most effective data mining methods and an approved technique to forecast survival of thyroid cancer patients [12]. Delen et al, used ANN, decision tree, and logistic regression to improve breast cancer forecast. They used the decision tree for extracting knowledge out of the data [13]. Landing et al, employed ANN and logistic regression to produce 5, 10, and 15 years forecasts of breast cancer patients. They used size of tumor, lymphatic node status, type of tissue, formation of tubolos, tumor necrosis and age as input variables and concluded that clustering trees and logistic regression are more effective for clinical interpretations [14]. Tolouie et al studied recurrence of breast cancer and supported vector machines (SVM) for forecasting recurrence of breast cancer, minimum error, and maximum accuracy [15].

In addition, ANN has been used for forecasting survival of esophageal cancer patients in [16], mortality rate among the liver hepatocellular carcinoma patient [17], and awareness among the liver cancer patients [18]. ANN outperformed statistical methods and TNM staging system in determining stage of different types of cancers [19].

Data mining and diabetes

Diabetes is a chronic and complicated disease with several symptoms. Industrialization of human life has brought increase in diabetes cases. About 200 million diabetics live in the world and 2 million of them are in Iran.

Miak et al, used card method to study the factors in development of diabetic symptoms [20]. Regression method was employed by Rolfing et al, to examine the relationship between blood sugar of diabetic I and HbAlc [21]. Big Hoan Cho et al, used SVM through feature section and visualization to find neuropathy among the diabetic [22].

ANN, decision tree, and logistic regression were compared regarding diagnosing diabetic or prediabetic patients among individuals with factor risk and the results showed that decision making tree had the highest susceptibility (75.13) and accuracy (77.87), and ANN had the lowest accuracy (73.23) [23].

Kim et al, used data mining (a priori algorithm) to survey the co-diseases and symptoms of diabetes and relationship between these, among 411414 patients [24]. Gregori et al, used data mining for monitoring diabetic patients [25]. Several studies have used different data mining algorithms to diagnose, manage and follow up diabetic patients. Ameri et al, used data mining algorithm (decision tree C5.0 and ANN) to classify diabetic patients based on their symptoms. They obtained the best results from the tree algorithm with accuracy of 89.06% and authenticity of model of 89.74% [26].

Data mining and renal diseases

Sepehri et al, employed decision tree to determine ureter stone treatment and argued that the model helped more patients to reach complete treatment [27]. Data mining techniques are mostly used for dialysis patients. Different available data mining techniques such as decision tree, fuzzy algorithm and so on are used in management and making decisions pertinent to dialysis patients and determining risk of cardiovascular diseases. In addition, these techniques are helpful in detecting early failure of venous fistul [28-30]. Decision tree has been showed to be more effective in follow up of the kidney implantation patients and the effects of risk

factors [31].

Data mining and cardiovascular diseases

Given the gravity of cardiovascular diseases -as the first cause of death in modern society- early diagnosis of the diseases is vital. Dehghani et al, used clustering data mining techniques to detect and forecast heart attacks [32]. Another study compared conventional data mining techniques such as decision-making tree, simple ANN Bayes, nearest neighbor K, and decision-making list and modern data mining tools such as weight association classifier (WAC) to achieve a proper algorithm to ensure accuracy of heart disease forecast. The results showed higher performance of simple ANN Bayes and decision tree [33]. Heart diseases analysis using evidence-based data mining techniques also known as Dampster-Shafer theory was also subject matter of another study. Austin et al, compared regression tree, ensemble-based methods, and conventional logistic regression to determine short-term (1 month) mortality rate among serious heart attack patients, and heart congestive failure. They concluded that boosted regression trees, trees and random forests outperformed conventional regression tree [34].

Data mining in health and hygiene

Data mining is also very useful in health and hygiene fields [35]. Electronic medical files can encompass plenty of data about symptoms, treatment, laboratory and medical results and so on. Valuable information can be extracted from these data [36-38]. In addition, the information can be helpful for hospital infection controls [39], ranking hospitals [39], and implementation of health services [40-41].

Conclusion

What mentioned above was few cases of data mining applications in providing health services to cancer, diabetic, kidney, and cardiovascular patients and health and hygiene services. Over the 20 years since the introduction of the concept, the applications have grown rapidly in medical and health fields. Apparently, data mining technique will extend to all fields of health and become one of the main decision making instruments in diagnosis, treatment, and health policy making. It is recommended therefore physicians, researchers, and decision makers put more emphasis on data mining concepts along with statistics and health system management techniques.

Acknowledgments

Author thank from staff of Industrial Intelligence Research Group.

Authors' Contribution

All authors had equal duties in this article.

Financial Disclosure

The authors declared no financial disclosure.

Funding/Support

This work did not receive any funding or support.

References

1- Adel A, Ahmadi P, Sebt M. Designing model for choosing human resources with data mining approach. Journal of Iranian Technology 2010; 2(4): 5. [In Persian]
2- Han J, Kamber M, Pei J. Data mining: concepts and techniques. 3rd ed. Philadelphia: Elsevier; 2011.
3- Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI magazine. 1996;17(3):37.
4- Mullins IM, Siadaty MS, Lyman J, Skully K, Miller WG, et al. Data mining and clinical data repositories: Insights from a 667,000 patient data set. Comput Biol Med. 2006 Dec;36(12):1351-77.
5- Fayyad U, Piatetsky-Shapiro G, Smyth P, Uthurusamy R. Advance in knowledge discovery and data mining. 01 February 1996.
6- Hayati Z, Sadeghi Mojarad M, Jafari N. Discovery of electronic information, track user's movement using association rules algorithm in data mining: a case study of the University Library website URL STS Australia. Ketabdari VA etelaresani. 1389(13):251-283. [In Persian].
7- Park JE. Parks textbook of preventive and social medicine, 18th edition, M/s Banarsidas Bhanot Publishers, India, 2005. p: 162-183.
8- Jekel J, Katz D, Elmore J. Epidemiology, biostatics, and preventive medicine. sec. ed. W.B. Saunders comp, 2001, p: 52-54.
9- Han J, Kamber M. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 2001.
10- Hand DJ. Statistics and data mining: Intersecting disciplines. ACM SIGKDD Explorations. 1999;1(1):16-19
11- Asadi N, Sadrodini M. Employing data mining to identify cancer risk factors and determine the optimal treatment in Namazi hospital cancer database. 16th Annual National Conference of Computer Society of Iran, 2010; Sharif University.
12- Jajroudi M, Baniasadi T, Kamkar L, Arbabi F, Sanei M, Ahmadzadeh M. Prediction of Survival in Thyroid Cancer Using Data Mining Technique. Technol Cancer Res Treat. 2014 Aug;13(4):353-9.
13- Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. J. Artificial Intelligence in Medicine. 2010;34:113-27.
14- Lundin M, Lundin J, Burke HB, Toikkanen S, Pylkkanen L, Joensuu H. Artificial neural networks applied to survival prediction in breast cancer. Oncology. 1999;57(4):281-6.
15- Tooloei A, Pourebrahimi A, Ebrahimi M et al. Using Data Mining Techniques for Prediction Breast Cancer Recurrence. Iranian Journal of Breast Disease. 2013;5(4):23-34.
16- Sato F, Shimada Y, Selaru FM, Shibata D, Maeda M, Watanabe G, et al. Prediction of survival in patients with esophageal carcinoma using artificial neural networks. Cancer. 2005;103(8):1596-605.
17- Chiu HC, Ho TW, Lee KT, Chen HY, Ho WH. Mortality predicted accuracy for hepatocellular carcinoma patients with hepatic resection using artificial neural network. Scientific World Journal. 2013 Apr 30;2013:201976.
18- Hanai T, Yatabe Y, Nakayama Y, Takahashi T, Honda H, Mitsudomi T, et al. Prognostic models in patients with non-small-cell lung cancer using artificial neural networks in comparison with logistic regression. Cancer Sci. 2003;94(5):473-7.
19- Burke HB, Goodman PH, Rosen DB, Henson DE, Weinstein JN, Harrell FE, et al. Artificial neural networks improve the accuracy of cancer survival prediction. Cancer. 1997;79(4):857-62.
20- Miyaki K. Takei I. Watanabe K. Nakashima H, Watanabe K, Omae K. Novel statistical classification model of type 2 diabetes mellitus patients for tailor-made prevention using data mining algorithm. J epidemiol. 2002;12(3): 243-8.
21- Rohlfing CL, Wiedmeyer HM, Little R, England JD, Tennill A, Goldstein DE. Defining the relationship between plasma glucose and HbA1c: analysis of glucose profiles and HbA1c in the Diabetes Control and Complications Trial. Diabetes Care. 2002;25(2):275-8.
22- Huang Y. McCullagh P. Black N. Harper R. Feature selection and classification model construction on type 2 diabetic patients' data. Artif intell med. 2007;41(3):251-62.
23- Meng XH, Huang YX, Rao DP, Zhang Q, Liu Q. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. Kaohsiung J Med Sci. 2013 Feb; 29(2):93-9.
24- Kim HS, Shin AM, Kim MK, Kim YN. Comorbidity study on type 2 diabetes mellitus using data mining. Korean J Intern Med. 2012 Jun;27(2):197-202.

25- Gregori D, Petrinco M, Bo S, Rosato R, Paqano E, Berchialla P, Merletti F. Using data mining techniques in monitoring diabetes care. The simpler the better?. J Med Syst. 2011;35(2):277-81.
26- Ameri H, Alizade S, Barzegari A. Knowledge extraction of diabetics' data by decision tree method. Journal of Health Administration. 2013:53(3):58-72. [In Persian]
27- Sepehri MM, Rahnama P, Shadpour P, Teimourpour B. A data mining based model for selecting type of treatment for kidney stone patients. Tehran University Medical Journal. 2009;67(6):421-7.
28- Rezapour M, Khavanin Zadeh M, Sepehri MM. Implementation of predictive data mining techniques for identifying risk factors of early AVF failure in hemodialysis patients. Comput Math Methods Med. 2013:830745.
29- Shah S, Kusiak A, Dixon B. Data mining in predicting survival of kidney dialysis patients, in proceedings of photonics west. Bios 2003, Bass, L.S. et al. (Eds), Lasers in Surgery: Advanced Characterization, Therapeutics, and Systems XIII, Vol. 4949, SPIE, Belingham, WA, January 2003, p:1-8.
30- Kusiak A, Bradley Dixonb, Shital Shaha. Predicting survival time for kidney dialysis patients: a data mining approach. Computers in Biology and Medicine; 2005:35(4):311-327.
31- Greco R, Papalia T, Lofaro D, Maestripieri S, Mancuso D, Bonofiqlio R. Decisional trees in renal transplant follow-up. Transplant Proc. 2010 ;42(4):1134-6
32- Dehghani T, Afshari Saleh M, Khalilzadeh M. A genetic K-means clustering algorithm for heart disease data. 5th Conference of Data Mining of Iran, 2011; Amirkabir University.
33- Zamanpoor S, Shamsi M. Assess and compare the accuracy of data mining algorithms to predict a heart disease. 4th Iranian Cnference on Electrical and electronics Engineering, 1391 Gonabad, Iran.
34- Austin PC, Lee DS, Steyerberg EW, Tu JV. Regression trees for predicting mortality in patients with cardiovascular disease: what improvement is achieved by using ensemble-based methods?. Biom J. 2012;54(5):657-73.
35- Koh HC, Tan G. Data Mining Application in Healthcare. J Healthc Inf Manag. 2005 Spring;19(2):64-72.
36- Balib RK. Clinical Knowledge Management: Opportunities and Challenges. Hershey: Idea Group Inc (IGI); 2005.
37- Trifirò G, Pariente A, Coloma PM, Kors JA, Polimeni G, Miremont-Salame G, et al. Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor?. Pharmacoepidemiol Drug Saf. 2009;18(12):1176-84.
38- Warner JL, Zollanvari A, Ding Q, Zhang P, Snyder GM, Alterovitz G.Temporal phenome analysis of a large electronic health record cohort enables identification of hospital-acquired complications. J Am Med Inform Assoc. 2013 Des;20(e2):e281-7
39- Obenshain MK. Application of data mining techniques to healthcare data. Infect Control Hosp Epidemiol 2004;25(8): 690-5.
40- Rogers G, Joyner E. Mining Your Data for Healthcare Quality Improvement [Online]. 2011 [cited 2011 Aug 8]; Available from: URL: http://www2.sas.com/proceedings/sugi22/EMERGING/PAPER139.PDF
41- Cios KJ. From the guest editor medical data mining: knowledge discovery in a clinical data warehouse. Engineering in Medicine and Biology Magazine, IEEE. 2000;19(4):15-6.